1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January, 1994	Fourth Quarterly R&D Status Rpt.

4. TITLE AND SUBTITLE

A Redesigned Isis and Meta System under Mach

5. FUNDING NUMBERS

N00014-92-J-1866

6. AUTHOR(S)

Professor Kenneth Birman, Professor Keith Marzullo

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Kenneth Birman, Associate Professor
Department of Computer Science
cornell University

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

DARPA/ONR

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE
DISTRIBUTION UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

14. SUBJECT TERMS	15. NUMBER OF PAGES 10
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

Redesigned Isis and META System under Mach

Fourth Quarterly R & D Status Report
Second Semi-Annual R & D Status Report
Jan. 1, 1994

Prof. Kenneth P. Birman
Department of Computer Science
Cornell University, Ithaca New York
607-255-9199

Prof. Keith Marzullo
Dept. of Computer Science
U.C. San Diego, San Diego, California

This work was sponsored by the Defense Advanced Research Projects Agency (DoD), under contract N0001492J1866 issued by the Office of Naval Research.

Personnel

- Academic Staff:
  - Prof. Kenneth P. Birman (Cornell), Principle-Investigator
  - Prof. Keith Marzullo (U.C. San Diego), Co-Investigator
  - Dr. Robert Cooper, Research Associate
  - Dr. Robbert van Renesse, Research Associate
- Graduate Students:
  - Lorenzo Alvisi (Marzullo)
  - David Cooper (Birman)
  - Brad Glade (Birman)
  - Guerney Hunt (Birman)
  - Sophia Georgiakaki (Birman)
  - Katie Guo (Birman)
  - Neil Jain (Birman)
  - David Karr (Cooper)
  - Michael Kalantar (Birman)
  - Laura Sabel (Marzullo)
  - Alexey Windbuhl (Birman)
  - Devin Barnhart
  - Matthew Clegg
  - Ida Szafranska (Marzullo)

The Horus project

This quarterly status report covers activities of the Horus project during the fourth quarter of 1993. This is also our second semi-annual progress report under the present ARPA grant. Because these status reports are intended to be brief and our proposal was recently funded, we assume that the reader has some background regarding the goals and status of our effort, and focus instead on technical accomplishments during the report period and goals for the next three months. Readers unfamiliar with our work could start by reading some of the papers cited below. The Isis overview that appeared in Communications of the ACM in December 1993 gives a good general picture of our past work.

This report uses the new ONR reporting style. It discusses *recent progress, transitions,* and *recent publications.*

Progress

With Marzullo now solidly installed at U.C. San Diego, our work has two main tracks. The larger of these is the continuing Horus development effort at Cornell University, which has now resulted in a working system that others have begun to experiment with. Near term goals with Horus are focused on developing an appropriate API for users and embedding the system in settings with minimal need for general purpose operating systems features.

The San Diego effort focuses on a real-time technology integrated with Horus, called the Corto subsystem.

In addition to these activities, we are involved in two significant collaborations, one with the Transis project (located at Hebrew University in Jerusalem), and one with the Delta-4 group (located at INESC in Portugal).

The major accomplishments of this report period are as follows:

- We have continued to extend and build upon the first version of the Horus system, following a path similar to the one used in developing the older ISIS toolkit. Predictions of a 10- to 100-fold performance improvement appear to be justified. With the recent delivery of a new release of X-kernel under native Mach, we are resuming work on a port of our software that would run under Mach.

- Working with Hebrew University's Transis group, we developed a new approach to tolerating partitions and have integrated the necessary mechanisms into Horus. With this code in place, it is possible to develop applications that continue operating even during network partition failures and that automatically "heal themselves" upon re-establishing communication.

- Also working with Hebrew University's Transis group, we developed a way to present systems like Horus and Transis through the UNIX socket interface, or the Mach ports interface. Our approach has the benefit of not requiring changes to UNIX or Mach, and in fact we plan to run Horus on the Mach microkernel (with none of the remainder of the operating system) as an experiment during 1994.

- We developed new and much improved flow control algorithms for Isis and for Horus, which also permit expanded use of hardware multicast when that feature is available. Performance

exceeds that reported for any other general purpose UNIX-based multicast algorithm, although it trails the performance obtained in the Amoeba project using "raw" hardware (a special purpose device driver) and in the Transis project, which does not support any sort of toolkit or application development environment.

- We developed a new "cellular" approach to presenting Horus and Isis, as well as other systems. Briefly, the idea starts with the recognition that these technologies can only enhance reliability, not protect completely against catastrophic failure. Indeed, they can even introduce new types of distributed failures that originate in bugs or deadlocks in our own code, although, obviously, we do everything we can to minimize this! A *cellular* presentation of a system divides the system into multiple cells, using very high performance gateways for inter-cell communication. Our cellular approach is for process-group systems, and our gateways can be used between Horus cells or Isis cells, but can also be used to connect Horus to Isis, and indeed to connect either of these systems to Transis or Delta-4. We are currently writing a paper on this approach. Cells are also useful in multi-level security domains. An example of how the architecture will be used arises in the French Air Traffic Control System. In this system, each position (3 consoles operated by a team of flight controllers) would be a single cell; the system-wide management and monitoring system would be another cell. Such an architecture scales very well, and it weakens the dependency between software on different cells. With care, it should be possible to design this system to tolerate even a completely arbitrary failure in one cell – other cells would just keep running normally, perhaps spooling some data for transmission to the failed cell when it recovers. We view this as a big conceptual advance for the project and one that is likely to have significant impact on the feasibility of using Horus and Isis in very large settings or in very critical ones.

- We are making good progress towards the design and implementation of Corto, the real-time layer. The lowest layers of this system consist of three major components: a clock synchronization service, a real-time multicast transport service, and a network scheduling layer that synchronizes with the multicast transport layer and the POSIX-compliant threads scheduling facility. This software is running on top of Lynx, a commercial real-time kernel, but we are designing the system to also run on top of any of the real-time Mach systems as they become available (we expect to migrate to RT-Mach by the beginning of the Summer).

  Early in our design process, it became clear that a TDMA-based approach (such as used by Mars, an important and highly influencial system recently built by Hermann Kopetz at the Universit of Vienna) would provide the best performance and schedulability as compared to either sequencing site protocols (such as Horus or Amoeba) or train-based protocols (such as Transis and Total). It also was clear that the optimization provided by CBCAST in the asynchronous Horus suite of protocols is not applicable to in a real-time domain. In addition, we found that an end-to-end argument could be used to double the bandwidth (at a cost of increased maximum latency) of a TDMA-based approach. Our resulting real-time multicast transport protocol is fast, dependable in the face of processor crashes and various ommission failures, and is highly portable. We are now working on how to make the assignment of TDMA slots dynamic in order to guarantee hard real-time delivery requirements in the face of changing membership and changing mode. We are using a simple approach that will not be as efficient as that provided by Mars but will be much more adaptable. Our hope is to eventually adapt the on-line scheduling techniques developed by Christos Papadimitriou of U.C. San Diego to slot scheduling.

We have had the prototype protocols running for approximately three months now, and are mainly attacking problems with operating system, hardware and scheduling constraints. For example, we wish to have Corto be as portable as possible across commericial POSIX-compliant Unix platforms. Hence, we have built the real-time transport service on top of UDP, using a single Ethernet and using the standard Unix sleep library routine. This approach places a lower bound on the TDMA slot of 10 msec (unless one staggers the sleep interrupts. We have done so, thereby reducing the slot to 2.5 msec for a 4-processor system, but the approach obviously doesn't scale well). In any case, the prototype maintains clock synchronization of approximately 300 microseconds with resynchronization occuring approximately twice a minute, and message stabiiity occurs in two TDMA slots.

- We completed several papers (see below under "publications") and a book.
- Continuation of research ties with other laboratories, including the Los Alamos Advanced Computing Laboratory (which focuses on supercomputing), the Isreali Transis project, Portugal's INESC research laboratory (known for its work on realtime communication), and with Mach-related research efforts at the Open Software Foundation, Carnegie Mellon, and University of Arizona.
- We continued the implementation of the security architecture for Horus, by extending the existing code to include a secure name service.
- We have continued our new effort to explore specialized implementations of Horus for parallel processors and for ATM networks. This work is motivated by the impressive results of Berkeley's Split/C and Active Messages research, demonstrating that asynchronous communication can lead to tremendous performance gains on the most important emerging parallel processors. In a very exciting development, the main developer of the Active Messages system (Thorsten von Eiken) joined our project during the fall, as a faculty member in the Cornell Dept. of Computer Science. Brian Smith, who has worked on multimedia file servers, will join us shortly. This has created the critical mass for a push that will move Horus onto highly parallel platforms, and onto advanced high performance communication environments. We want to build our protocols in ways that exploit the hardware fully and minimize unnecessary work in software – work needed on networks but not on closely coupled machines. We are very pleased with this new direction.

Transitions

Our project is perhaps unique among distributed systems efforts in the United States in the degree of success we have had with technology transfer.

Technology transfer

During the fourth quarter of 1993, Stratus Computer Inc. of Boston, a company specialized in availability technologies, acquired Isis Distributed Systems, our spin-off company that has focused on commercializing Isis. The acquisition came at a time of a great success for Isis, which has been selected for use in settings like the New York Stock Exchange, the French Air Traffic Control System, the Swiss Electronic Bourse (ATB/EBS), Iridium, Sematech's factory floor system, and a great number of other high visibility, demanding applications. These include a number of U.S. government and military applications, of which the Hiper-D project (follow-on to AEGIS) is most visible, but extending to at least a dozen similar efforts in every agency of the miltary and government.

Stratus is firmly committed to availability, through software and hardware, and views Isis as the center of its future strategy in building continuously available computing systems and highly available distributed software. The company has stated emphatically that it will continue to port Isis to a wide variety of vendor platforms, including both UNIX systems and non-UNIX environments. The company will also be exploring ways to exploiting emerging hardware such as ATM technology that has the potential to dovetail with Horus, and has obtained rights to commercialize Horus through Cornell University. The committment to a heterogeneous presentation of Isis and Horus has been repeatedly stressed by Stratus, which intends to port Isis and Horus to a wider range of platforms (notably, PC's) while maintaining the technology on the current platforms (mostly UNIX workstations and VMS).

Through this development, Isis and Horus are clearly entering the mainstream and will have greatly enhanced impact on the economic mission of the country, just at a time when the demand for reliability in "data highway" applications is becoming acute. We believe that this is a great success story for ARPA and ONR: a proof that the decade-long investment by ARPA in this technology area has created the basis for a new industry that has become self-sustaining and accepted.

The Stratus acquisition will cause some reorganization within the Horus effort at Cornell, but nothing drastic is expected to change. Birman will consult for Stratus in the role of Chief Scientist for the Isis effort, but this is recognized as a part-time job, and is not expected to create more load than Birman's work as President of Isis over the past few years. Birman will continue to head the Horus project, where the focus now is on arriving at a mature system that Stratus can pick up, while also striking in a new direction that would explore ATM networking and parallel computing.

Cooper will be leaving Cornell to head the technology development group within Isis at Stratus, but will remain in Ithaca and will continue to work closely with us. We currently plan to fill his position with a post-doctoral student. Van Renesse will remain at Cornell, but will consult for Stratus to assist in technology transition for Horus.

Stratus is committed to maintaining access to Isis and Horus for research users in academic settings and other settings. The structure appears to be an ideal route for transition of current and future work by our research effort.

Hiper-D effort

Isis continues to work closely with the HiperD program, which is now being moved to the Navy and will become an advanced prototyping effort under the overall AEGIS R&D effort. This work seems to be moving forward rapidly, and has adopted a reasonable compromise between needing to use robust existing technology (Isis, Mach) and wanting to exploit emerging platforms like the Paragon. Isis Distributed Systems will maintain a significant effort in this area, and will continue to provide any necessary support to the HiperD developers at JH/APL and elsewhere.

Collaborations

As noted above, Horus has excited wide interest in the research and advanced development community. We maintain close ties to dozens of other efforts, and are sharing technology with several national laboratories, supercomputing projects at Los Alamos Laboratories, Sandia, and NASA JPL, and are exploring ties with a number of commercial prototyping efforts.

=== ADMINISTRATIVE DATA ===

1 ARPA ORDER NUMBER: 9247

2 CONTRACT/GRANT NUMBER: N00014-92-J-1866=20

3 AGENT: ONR

4 CONTRACT TITLE: A Redesigned ISIS and Meta System under Mach=20

5 CONTRACTOR/ORGANIZATION: Cornell University=20

6 SUBCONTRACTORS: Univ of Calif,San Diego $230,954

7 PRINCIPAL INVESTIGATORS:

Kenneth Birman
Cornell Univ
4105A Upson Hall
Ithaca, NY 14853
Phone:607-255-9199
Fax: 607-255-4428
Email: ken@cs.cornell.edu

Keith Marzullo
Univ of Calif, San Diego
La Jolla, CA 92093
Phone:(619) 534-3729
Fax: (619) 534-7029
Email: marzullo@cs.ucsd.edu

8 ACTUAL START DATE: Sept 30,1992

9 EXPECTED END DATE: December 30,1995

10 FUNDING PROFILE: @ 12/31/93

10.1 Current contract:

FY93	FY94	FY95	TOTAL
$1,281,331	$900,000	$956,187	$3,137,518

10.2 Options (one line for each)

NA

10.3 Total funds provided to date. $1,281,331

10.4 Actual (est) funds expended through December 31,1993: $1,185,700

10.5 Date current funding will be expended: January 31,1994

10.6 Funds required in FY94 by quarter through 12/31/94:

1/94-3/94	4/94-6/94	7/94-9/94	10/94-12/94
$225,000	$225,000	$225,000	$225,000

10.7 Date. 1/11/94

11 ANYTHING ELSE YOU NEED: N/A

Fourth Budget Statement

a.	ARPA Order Number:	9247
b.	Contract Number:	N00014-92-J-1866
c.	Agent:	ONR
d.	Contract Title:	A Redesigned ISIS and Meta System Under Mach
e.	Organization:	Cornell University
f.	PIs:	Kenneth P. Birman and Keith Marzullo
g.	Actual Start Date:	9/30/92
h.	Expected End Date:	12/30/95
i.	Expected End Date if Options Exercised:	N/A
j.	Total Price:	$3,137,518
k.	Spending Authority Provided So Far:	$1,281,331
l.	Expenditures through 9/93	$947,400
m.	Date When These Funds Will Be Fully Expended:	1/31/94
n.	Additional Funds Expected Per Contract (by FY):	FY94 $900,000 FY95 $956,187

Publications

Below, we reproduce a list of recent publications by the effort. A good general review of the project is the article that appeared in the December issue of Communications of the ACM. We have also just completed a book that will be published by IEEE Press, and collects the most important papers Isis papers together with about 50% previously unpublished material, as a single volume. The book will appear early in 1994.

PUBLICATIONS LIST

ISIS Activity

- Keith Marzullo and Ida Szafranska. Monitoring and Controlling Distributed Applications using Lomita. IEEE First International Workshop on Systems Management, 14-16 April 1993.
- Ken Birman. A Response to Cheriton and Skeen's Criticism of Causal and Totally Ordered Communication. Published in Operating Systems ReviewJanuary, 1994.
- Lorenzo Alvisi, Bruce Hoppe and Keith Marzullo. Nonblocking and Orphan-Free Message Logging Protocols. Accepted for presentation at FTCS.
- Ozalp Babaoglu, Keith Marzullo and Fred B. Schneider. Priority Inversion and its Prevention. Accepted for publication in Journal of Real-Time Systems, volume 5, number 4.
- The Process Group Approach to Reliable Distributed Computing. Kenneth P. Birman. *Communications of the ACM*, 36:12 (Dec. 1993).
- André Schiper, Aleta Ricciardi and Kenneth Birman. Virtually-Synchronous Communication Based on a Weak Failure Suspector. June 22-24, 1993, Toulouse, France.
- Robbert van Renesse. Why Bother with CATOCS? Published in Operating Systems Review, January, 1994, Vol. 28, No 1.
- Robert Cooper. Experience with Causally and Totally Ordered Communication Support. Published in Operating Systems Review, January, 1994, Vol 28, No 1.
- Robbert van Rennesse and Dag Johansen. Distributed Systems in Perspective. Published in Distributed Open Systems.
- Kenneth Birman and Brad Glade. Consistent Failure Reporting in Reliable Communication Systems. Submitted October, 1993.
- Kenneth Birman, A book to be published in the IEEE Press in early 1994. This books collects the most important Isis papers along with previously unpublished material.
- Robbert van Renesse and Kenneth Birman. Fault-Tolerant Programming Using Process Groups, Published in Distributed Open Systems
- Robbert van Renesse and Dag Johansen. Software Structures for Supporting Distributed Computing. Published in Distributed Open Systems.

- Keith Marzullo and M.D. Wood. Tools for Monitoring and Controlling Distributed Applications. Published in Distributed Open Systems.
- Bradford B. Glade, Kenneth P. Birman, Robert Cooper, and Robbert van Renesse. Light-Weight Process Groups in the ISIS System. Published in Distributed Systems Engineering Journal, July 1993.